



Universiteit  
Leiden  
Governance and Global Affairs

udp UNIVERSIDAD  
DIEGO PORTALES



# Benchmarking LLMs in Political Content Text-Annotation

Proof-of-Concept with Toxicity and Incivility Data

---

**Bastián González-Bustamante**

Leiden University & Universidad Diego Portales

✉ [b.a.gonzalez.bustamante@fgga.leidenuniv.nl](mailto:b.a.gonzalez.bustamante@fgga.leidenuniv.nl)

Presentation at the IV Training Data Lab Workshop

November 15, 2024

# Introduction

---

# The Challenge of Online Toxicity

**Digital platforms and incivility.** The rise of digital platforms has seen a surge in incivility and toxicity in online interactions, particularly in political discussions.

**Need for automated solutions.** The sheer volume of online content makes manual annotation impractical, necessitating automated solutions for analysing toxicity and incivility.

**LLMs as a potential solution.** LLMs offer a promising avenue for automating the annotation process, potentially revolutionising how we study online political discourse.



# Introducing LLMs to the Field

- 📄 **Traditional text analysis.** Traditional approaches to analysing text data in social sciences often relied on manual coding, dictionaries, or topic modelling, which can be time-consuming and resource-intensive.
- 🌐 **The power of LLMs.** LLMs, with their zero- and few-shot learning capabilities, offer a powerful alternative, enabling researchers to process and analyse vast amounts of text data efficiently.
- 🏛️ **Transforming research.** This shift towards AI-powered analysis has the potential to transform computational social science research, particularly in understanding online toxicity and incivility.

# OpenAI's GPTs: A Closer Look

## GPT and research

Researchers have increasingly utilised OpenAI's GPT models through their **API for various tasks**, including text annotation and classification.

## Advantages of GPT models

GPT models offer several advantages, including **ease of use**, access to substantial computational resources, and **impressive performance** in many tasks.

## Concerns and limitations

However, concerns remain regarding **reproducibility**, data privacy, and the potential for bias in these proprietary models.

# The Rise of Open-Source LLMs

## Addressing the limitations

Open-source LLMs have emerged as a response to concerns surrounding proprietary models, emphasising transparency, **reproducibility**, and community-driven development.

## Performance and potential

Recent studies have shown that open-source LLMs can match or **even surpass** the performance of proprietary models in specific text annotation tasks, particularly in zero-shot classification.

## Cost-effective alternatives

Some open-source “Small Language Models” (SLMs) offer comparable performance with significantly **reduced computational requirements** and costs.

## Data and Methods

---

# Data and Gold Standard

## Novel dataset on protest events

- Flag of Argentina: Protests against the coronavirus and judicial reform (August 2020) ✓
- Flag of Chile: Social outburst in Chile (October 2019) ✓
- Flag of Brazil: Protests against education budget cuts (May 2019) ✗

~3.5M messages in total\*

\* ~5M including Brazil

## LLMs zero-shot classification

- Random, balanced sample of 1,000 messages

## Open-source LLMs deployed locally

- Minimum temperature to ensure reproducibility\*\*
  - 💻 Intel Core i9-14900K CPU
  - 💻 NVIDIA GeForce RTX 4070 Super Windforce OC 12GB GPU
  - 💻 64 GB of RAM (DDR5-4800)
  - 🐧 Windows Subsystem for Linux v2.1.5.0 and Ollama v0.1.44 and v0.3.10

\*\* We ran some temperature experiments to test models' hallucinations and reproducibility.

# Gold Standard

[README](#) [Code of conduct](#) [CC-BY-4.0 license](#) [GPL-3.0 license](#)

---

## gold-standard-toxicity

Gold Standard for Toxicity and Incivility Project  
Annotated Data in Spanish for Toxicity and Insults in Digital Social Networks

coverage 100% version v0.5.3 repo status Active DOI 10.5281/zenodo.12574288 license GNU GPLv3 license CC BY 4.0  
made with [Raspberry Pi 5](#) supported by [tDL](#)

### Overview

This repository contains data sets and materials for a gold standard elaboration on toxicity and incivility in the digital sphere based on human coding to benchmark algorithmic classification tasks with transformers and LLMs. The labelling progress is indicated in the coverage badge above.

We are labelling two samples of novel datasets of political digital interactions on Twitter (rebranded as X). The first set comprises almost 5 million data points from three Latin American protest events: (a) protests against the coronavirus and judicial reform measures in Argentina during August 2020; (b) protests against education budget cuts in Brazil in May 2019; and (c) the social outburst in Chile stemming from protests against the underground fare hike in October 2019. We are focusing on interactions in Spanish to elaborate a gold standard for digital interactions in this language, therefore, we prioritise Argentinian and Chilean data. The second set contains more than 31 million messages and more than 9 million interactions between 2010 and 2022, covering the election of members of the first Constitutional Convention in Chile, the drafting process and the referendum in which the proposal was rejected.

This project is generously funded by the OpenAI Academic Programme, 2024 FAE-UDP Research Grant, and partially by the St Hilda's College Muriel Wise Fund at the University of Oxford. The [Training Data Lab](#) research group also logistically supports this project.

The sample was manually annotated by a team of five high-skilled human coders on Labelbox.

Each message was hand-coded by two different coders of the team, therefore, we have **two ground-truth labels for each sample observation**.

GitHub repository.

# Gold Standard

We used a binary codification performed using the core definitions of Perspective for toxicity as follows:

## Human annotation

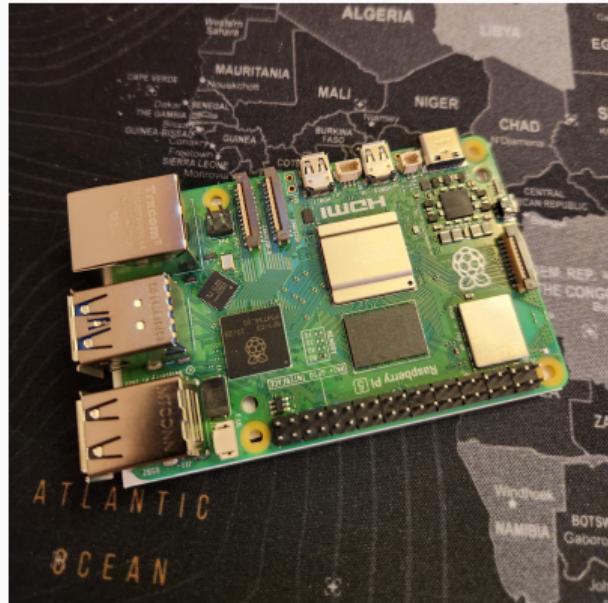
Classify the category of the comment as either **TOXIC** or **NONTOXIC**. **TOXIC**: Rude, disrespectful, or unreasonable comments that are likely to make someone leave the discussion or stop sharing their perspective. **NONTOXIC**: Civil or nice comments that are unlikely to discourage conversation.

- Krippendorff's  $\alpha$  (bootstrap 1,000x) 0.944 (95% CI [0.919, 0.961]).
- The process involved 7.2 hours of annotation and 1.2 hours of revision.

# Perspective API Classification on Raspberry Pi 5

First, we ran Perspective API, a distilled BERT developed by Jigsaw and Google. The classification implied **2,411 hours** of computing (**entire dataset, not only the sample**).

The process was deployed on a **Raspberry Pi 5**, a small computer board with an ARM CPU and meagre power consumption. Our footprint using Raspberry Pi 5 was 6.51 kg of CO<sub>2</sub>, which implied a **reduction of 96% in emissions** that we would have generated running the tasks on an average cloud computing service.



\* Picture of the Raspberry Pi 5 used for classification.

# LLMs for Text-Annotation: Prompt Strategy

Our prompt strategy was the very **same used on the human annotation** based on the core definitions of Perspective and Google for toxicity with the following system message:

## System message

Classify the category of the comment as either **TOXIC** or **NONTOXIC**. **TOXIC**: Rude, disrespectful, or unreasonable comments that are likely to make someone leave the discussion or stop sharing their perspective. **NONTOXIC**: Civil or nice comments that are unlikely to discourage conversation-

Along with providing texts of our balanced sample, we also listed the categories for the task as follows: **“Respond with only the category (TOXIC or NONTOXIC). Do not provide any additional analysis or explanation”**.

## Results

---

# Running Cycle 1: Baseline



Benchmarking LLMs in Political  
Content Text-Annotation

\* o1-preview o1-mini, released recently  
on September 12, 2024, are not fully  
available to all API users.

## OpenAI's GPTs\*

GPT-4o (2024-05-13), GPT-4o mini, GPT-4,  
GPT-4 Turbo and GPT-3.5 Turbo

## SOTA open-source LLMs

Llama 3.1 (8B), Hermes 3 (8B), Gemma 2 (9,  
27B) and Mistral NeMo (12B)

## Slightly outdated open-source LLMs

Nous Hermes 2 Mixtral (47B), Nous Hermes 2  
(11B), Mistral OpenOrca (7B), Orca 2 (7B) and  
Aya (35B)

## Benchmarking Cycle 1: Baseline

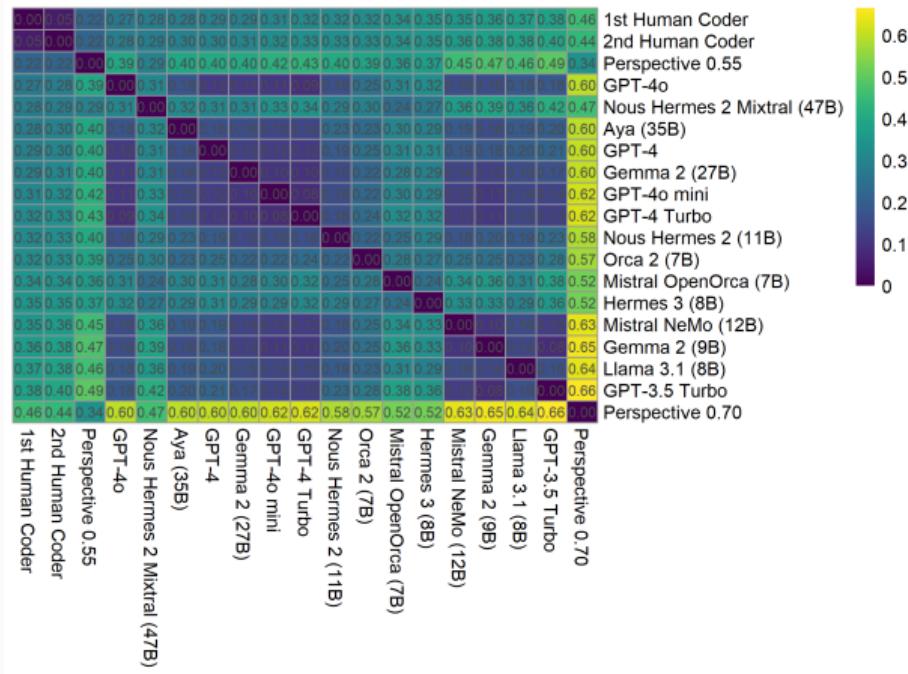
Model	Accuracy	Precision	Recall	F1-Score
Perspective 0.55	0.882	0.975	0.800	0.879
GPT-4o (2024-05-13)	0.804	0.735	0.991	0.844
Nous Hermes 2 Mixtral (47B)	0.829	0.859	0.813	0.835
Aya (35B)	0.793	0.727	0.979	0.835
GPT-4 (0613)	0.793	0.737	0.953	0.831
Gemma 2 (27B)	0.785	0.719	0.979	0.830
GPT-4o mini (2024-07-18)	0.761	0.695	0.985	0.815
GPT-4 Turbo (2024-04-09)	0.757	0.690	0.989	0.813
Nous Hermes 2 (11B)	0.772	0.727	0.918	0.811
Orca 2 (7B)	0.773	0.740	0.888	0.807
Mistral OpenOrca (7B)	0.777	0.790	0.794	0.792
Hermes 3 (8B)	0.770	0.770	0.811	0.790
Mistral NeMo (12B)	0.717	0.659	0.976	0.786
Gemma 2 (9B)	0.697	0.639	0.993	0.778
Llama 3.1 (8B)	0.706	0.659	0.931	0.772
GPT-3.5 Turbo (0125)	0.667	0.616	0.998	0.762
Perspective 0.70	0.756	1.000	0.543	0.704

F1-Score under 0.80

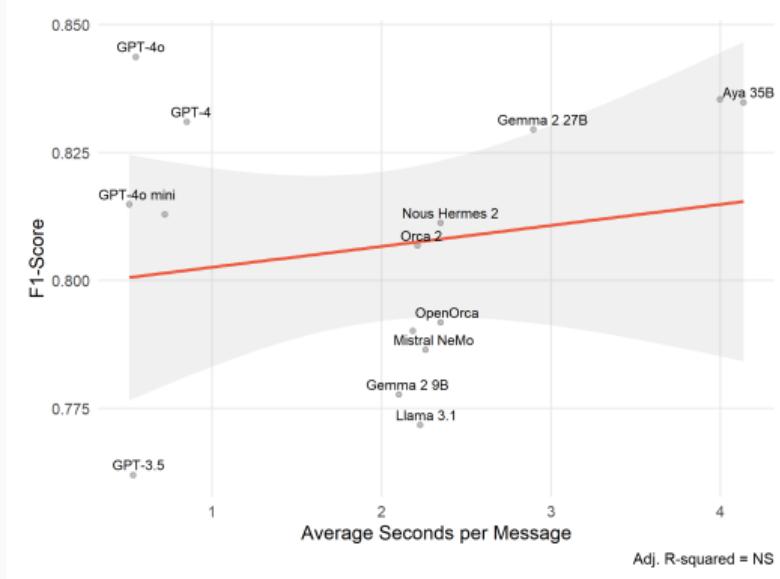
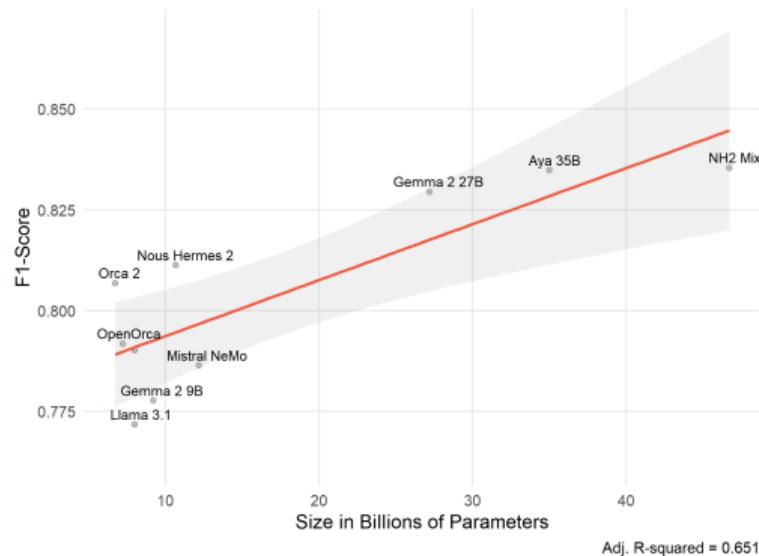
## Jaccard Distance between Gold Standard and Classifiers

With an intermediate performance, **Mistral**, **OpenOrca** and **Hermes 3**, both with parameters between 7 and 8B, show indexes around 0.30, therefore, they are not highly similar to other models. This could open avenues for ensemble annotations or stacking.

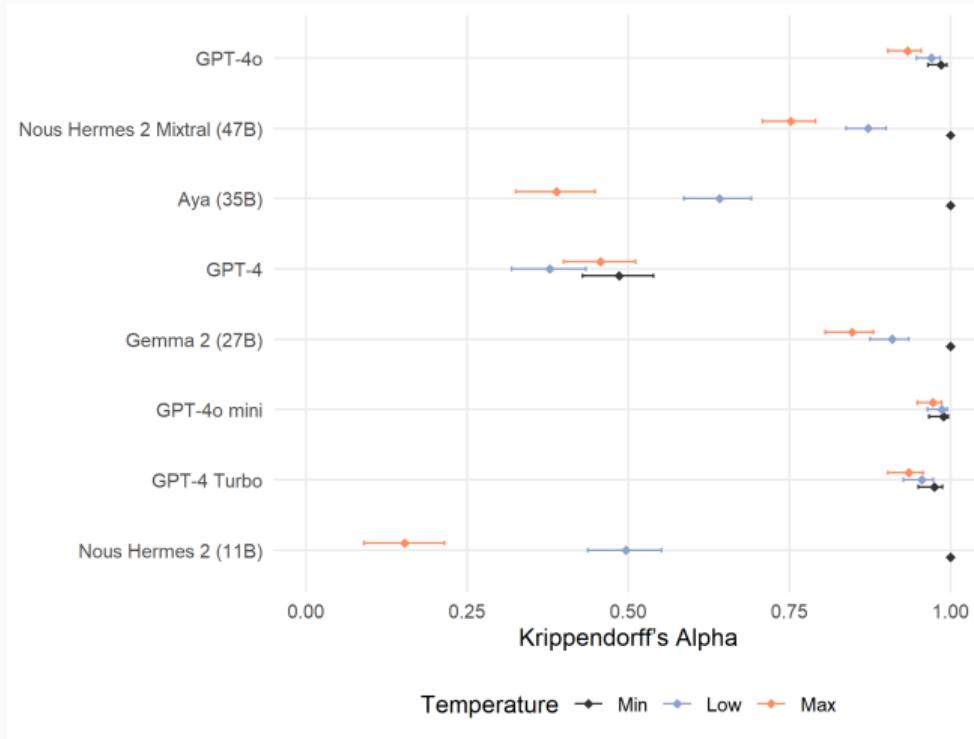
The same applies to Nous Hermes 2 Mixtral, which showed one of the highest performances.



# Performance, Parameters and Computing Time



# Computing Time vs Reproducibility



## ⌚ Time per obs.

Human coder: 15.125s

Perspective API: 1.175s

OpenAI's GPTs: 0.633s 🏆

Open-source LLMs: 2.672s

## 🐢 How much faster than humans?

Open-source LLMs ~6x

Perspective API ~13x

OpenAI's GPTs ~24x 🏆

# Running Cycle 2: ML and Transformers

## ML classification models

Naïve Bayes, SVM kernel-based linear, Random Forest with 100 and 500 and XGBoost

## Hyperparameters tuning

Additive smoothing, regularisation, limiting depth to control for overfitting, internal nodes and learning rate

## Fine-tuned BERT-ish models

ft RoBERTa (base on ~58M tweets)

ft XML-RoBERTa (base on 2.5TB multilingual data)

ft BERT Spanish (base BERT model for Spanish)



## Benchmarking Cycle 2: ML and Transformers (Top-15)

Model	Accuracy	Precision	Recall	F1-Score
Perspective 0.55	0.882	0.975	0.800	0.879
ft BERT Spanish*	0.835	0.816	0.886	0.849
GPT-4o (2024-05-13)	0.804	0.735	0.991	0.844
Nous Hermes 2 Mixtral (47B)	0.829	0.859	0.813	0.835
Aya (35B)	0.793	0.727	0.979	0.835
GPT-4 (0613)	0.793	0.737	0.953	0.831
Gemma 2 (27B)	0.785	0.719	0.979	0.830
GPT-4o mini (2024-07-18)	0.761	0.695	0.985	0.815
GPT-4 Turbo (2024-04-09)	0.757	0.690	0.989	0.813
Nous Hermes 2 (11B)	0.772	0.727	0.918	0.811
Orca 2 (7B)	0.773	0.740	0.888	0.807
Mistral OpenOrca (7B)	0.777	0.790	0.794	0.792
Naïve Bayes*	0.730	0.708	0.895	0.791
Hermes 3 (8B)	0.770	0.770	0.811	0.790
ft XML-roBERTa*	0.740	0.688	0.924	0.789

Rookie in the cycle; F1-Score under 0.80; \* 20% validation set, expected decrease in pseudo-test set

# Running Cycle 3: SOTA LLMs



## ☁️ OpenAI's GPTs

GPT-4o (2024-08-06)

## ☰ SOTA open-source LLMs

Llama 3.1 (70, 405B), Hermes 3 (70B), Aya  
Expanse (8, 32B), Qwen 2.5 (7, 14, 32, 72B),  
Mistral Small (22B) and Solar Pro (22B)

## ☰ Additional thresholds for Perspective

Perspective 0.60 and 0.80

## Benchmarking Cycle 3: SOTA LLMs (Top-15)

Model	Accuracy	Precision	Recall	F1-Score
Perspective 0.55	0.882	0.975	0.800	0.879
Qwen 2.5 (32B)	0.823	0.763	0.970	0.854
Perspective 0.60	0.862	0.995	0.745	0.852
GPT-4o (2024-05-13)	0.804	0.735	0.991	0.844
GPT-4o (2024-08-06)	0.802	0.735	0.985	0.842
Qwen 2.5 (72B)	0.804	0.741	0.972	0.841
Aya Expanse (32B)	0.804	0.748	0.955	0.839
Llama 3.1 (405B)	0.840	0.912	0.775	0.838
Hermes 3 (70B)	0.808	0.769	0.916	0.836
Nous Hermes 2 Mixtral (47B)	0.829	0.859	0.813	0.835
Aya (35B)	0.793	0.727	0.979	0.835
GPT-4 (0613)	0.793	0.737	0.953	0.831
Gemma 2 (27B)	0.785	0.719	0.979	0.830
Qwen 2.5 (14B)	0.799	0.756	0.921	0.830
Qwen 2.5 (7B)	0.776	0.727	0.929	0.816

Rookie in the cycle

## Cycle 4: Fine-tuning LLMs

### ft GPT-4o

- Base GPT-4o (2024-08-06)
- ft AOnla1B3
- 625,410 tokens
- 5 epochs
- Training loss 0.060
- API deployable

### ft Mistral (7B)

- Base Open Mistral (7B)
- ft 328207d0
- 1,310,720 tokens
- 8.563 epochs
- Training loss 0.081
- API deployable

### ft Mistral (12B)

- Base Mistral Large (12B)
- ft cbaac529
- 1,530,610 tokens
- 10 epochs
- Training loss 0.078
- API deployable

## Benchmarking Cycle 4: Fine-tuning LLMs (Top-15)

Model	Accuracy	Precision	Recall	F1-Score
ft GPT-4o**	0.874	0.882	0.931	0.906
ft Mistral (12B)**	0.847	0.849	0.931	0.888
Perspective 0.55	0.882	0.975	0.800	0.879
Qwen 2.5 (32B)	0.823	0.763	0.970	0.854
Perspective 0.60	0.862	0.995	0.745	0.852
GPT-4o (2024-05-13)	0.804	0.735	0.991	0.844
GPT-4o (2024-08-06)	0.802	0.735	0.985	0.842
Qwen 2.5 (72B)	0.804	0.741	0.972	0.841
Aya Expanse (32B)	0.804	0.748	0.955	0.839
Llama 3.1 (405B)	0.840	0.912	0.775	0.838
Hermes 3 (70B)	0.808	0.769	0.916	0.836
Nous Hermes 2 Mixtral (47B)	0.829	0.859	0.813	0.835
Aya (35B)	0.793	0.727	0.979	0.835
GPT-4 (0613)	0.793	0.737	0.953	0.831
Gemma 2 (27B)	0.785	0.719	0.979	0.830

Rookie in the cycle; \*\* pseudo-test set

# Extended Benchmarking Overview (Top-15)

Model	Accuracy	Precision	Recall	F1-Score	Deploy	CO <sub>2</sub>
ft GPT-4o**	0.874	0.882	0.931	0.906	API	☛
ft Mistral (12B)**	0.847	0.849	0.931	0.888	API	☛
Perspective 0.55	0.882	0.975	0.800	0.879	API-RPi5	leaf
Qwen 2.5 (32B)	0.823	0.763	0.970	0.854	Local	⚖️
Perspective 0.60	0.862	0.995	0.745	0.852	API-RPi5	leaf
GPT-4o (2024-05-13)	0.804	0.735	0.991	0.844	API	☛
GPT-4o (2024-08-06)	0.802	0.735	0.985	0.842	API	☛
Qwen 2.5 (72B)	0.804	0.741	0.972	0.841	Local	⚖️
Aya Expanse (32B)	0.804	0.748	0.955	0.839	Local	⚖️
Llama 3.1 (405B)	0.840	0.912	0.775	0.838	API	☛
Hermes 3 (70B)	0.808	0.769	0.916	0.836	Local	⚖️
Nous Hermes 2 Mixtral (47B)	0.829	0.859	0.813	0.835	Local	⚖️
Aya (35B)	0.793	0.727	0.979	0.835	Local	⚖️
GPT-4 (0613)	0.793	0.737	0.953	0.831	API	☛
Gemma 2 (27B)	0.785	0.719	0.979	0.830	Local	⚖️

Private-closed; Open-source; BERT-ish; \*\* psuedo-test set

# Error Rate vs 2nd Human Coder

**Predictive Performance Index (PPI).** Average performance metrics with equal weights.

**Geometric Mean (GPI).** Penalty on low values controlling volatility between metrics using the 4th root of the product of performance metrics.

Misclassification	PPI	GPI
Top-Tier Options		
ft GPT-4o vs 2nd Coder	7.50%	7.52%
ft Mistral (12B) vs 2nd Coder	9.45%	9.51%
Perspective 0.55 vs 2nd Coder	8.93%	9.14%
Mid-Range Option		
Orca 2 (7B) vs 2nd Coder	17.13%	17.31%
Sub-Optimal Options		
GPT-3.5 Turbo vs 2nd Coder	21.25%	22.56%
Perspective 0.80 vs 2nd Coder	32.68%	36.58%

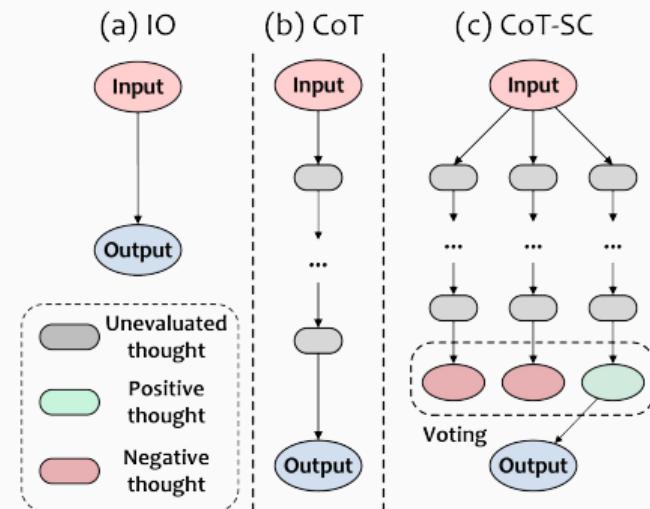
Private-closed; Open-source; BERT-ish

# Chain-of-Thought and Self-Consistency

Model	Prompting	Variation	
		PPI	GPI
GPT-4o	CoT-SC	3.84%	4.18%
GPT-4o	CoT	3.18%	3.47%
Qwen2.5 (32B)	CoT-SC*	1.38%	1.53%
Qwen2.5 (32B)	CoT*	1.29%	1.44%
ft GPT-4o	CoT-SC	0.17%	0.18%
ft GPT-4o	CoT	0.14%	0.16%
ft Mistral (12B)	CoT-SC*	-1.34%	-1.57%
ft Mistral (12B)	CoT*	-1.59%	-1.85%

Private-closed; Open-source

\*About 15% of outcome need more cleaning because of JSON issues.



Source: [Din et al. \(2024\)](#).

## Takeaways

---

## Takeaways

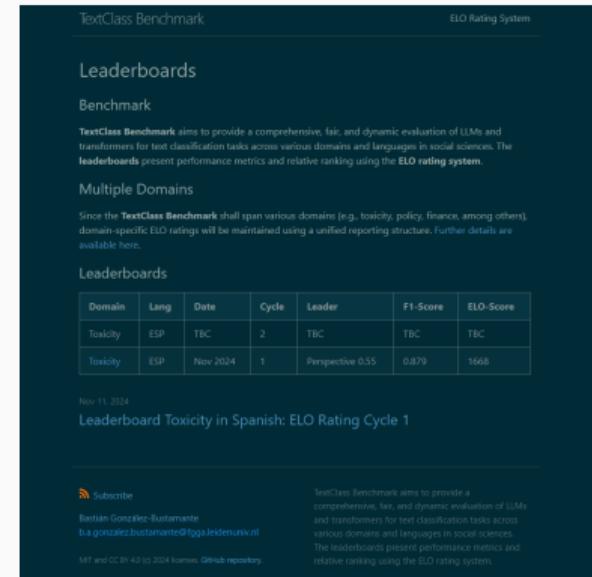
- 🏆 Fine-tuned GPT-4o and Mistral (12B) outperform other LLM's zero-shot classification. Perspective API using a laxer threshold still shows an outstanding performance.
- ⚠️ Classic machine learning approaches and even fine-tuned BERT-ish models are **not competitive against LLMs**.
- ⚠️ Sub-optimal choices could lead to **misclassification above 20%**.
- ☑️ Although GPTs tend to show not only excellent computing time but also overall good levels of reliability, **only open-source LLMs ensure full reproducibility** in the annotation.
- ☑️ **CoT shows a slight improvement**, ~1-2%, and it seems to work better in unfine-tuned models. The JSON structure could be problematic, and there is an important increase in computing time.

# TextClass Benchmark

**TextClass Benchmark** aims to provide a comprehensive, fair, and dynamic evaluation of LLMs and transformers for text classification tasks across various domains and languages in social sciences.

The leaderboards present performance metrics and relative ranking using the **ELO rating system**.

 [textclass-benchmark.com](https://textclass-benchmark.com)



The screenshot shows the TextClass Benchmark website's Leaderboards page. At the top, there are tabs for "TextClass Benchmark" and "ELO Rating System". Below the tabs, there are sections for "Leaderboards" and "Benchmark". The "Leaderboards" section contains a summary text about the project's aim to evaluate LLMs and transformers across various domains and languages in social sciences, mentioning the use of the ELO rating system for performance metrics and relative ranking. It also notes that the website spans multiple domains like toxicity, policy, finance, etc. The "Multiple Domains" section indicates that ELO ratings will be maintained using a unified reporting structure. Below these sections is a table titled "Leaderboards" showing performance metrics for different domains and languages. The table includes columns for Domain, Lang, Date, Cycle, Leader, F1-Score, and ELO-Score. One row is visible for Toxicity in Spanish (ESP) during Cycle 1, with a Perspective score of 0.55, an F1-Score of 0.479, and an ELO-Score of 1668. At the bottom of the page, there is a footer with links for "Subscribe" (RSS icon), contact information for Berlina González-Bustamante (berlina.gonzalez.bustamante@toga.istemi.mx), and copyright information (MIT and CC BY 4.0 © 2024 license, GitHub repository). A note in the footer states that the website aims to provide a comprehensive, fair, and dynamic evaluation of LLMs and transformers for text classification tasks across various domains and languages in social sciences, using the ELO rating system for performance metrics and relative ranking.

Domain	Lang	Date	Cycle	Leader	F1-Score	ELO-Score
Toxicity	ESP	TBC	2	TBC	TBC	TBC
Toxicity	ESP	Nov 2024	1	Perspective	0.55	0.479
						1668

Nov 11, 2024  
Leaderboard Toxicity in Spanish: ELO Rating Cycle 1

Subscribe

Berlina González-Bustamante  
[berlina.gonzalez.bustamante@toga.istemi.mx](mailto:berlina.gonzalez.bustamante@toga.istemi.mx)

MIT and CC BY 4.0 © 2024 license, GitHub repository

TextClass Benchmark aims to provide a comprehensive, fair, and dynamic evaluation of LLMs and transformers for text classification tasks across various domains and languages in social sciences. The leaderboards present performance metrics and relative ranking using the ELO rating system.

This work was supported by the OpenAI's Academic Programme and the Faculty of Administration and Economics at the Universidad Diego Portales, Chile. I also thank the Institute of Security and Global Affairs at Leiden University, Netherlands, and the Training Data Lab research group for their support.



Universiteit  
Leiden  
Institute of Security  
and Global Affairs



# Thank you very much!

Do you have any questions?

[b.a.gonzalez.bustamante@fgga.leidenuniv.nl](mailto:b.a.gonzalez.bustamante@fgga.leidenuniv.nl)

[bgonzalezbustamante.com](http://bgonzalezbustamante.com)